A 10-Gene Classifier for Distinguishing Head and Neck Squamous Cell Carcinoma and Lung Squamous Cell Carcinoma

Anil Vachani,[1] Michael Nebozhyn,[2] Sunil Singhal,[1] Linda Alila,[2] Elliot Wakeam,[1] Ruth Muschel,[7] Charles A. Powell,[3] Patrick Gaffney,[5] Bhuvanesh Singh,[4] Marcia S. Brose,[1] Leslie A. Litzky,[1] John Kucharczuk,[1] Larry R. Kaiser,[1] J. Stephen Marron,[6] Michael K. Showe,[2] Steven M. Albelda,[1] and Louise C. Showe[2]

Abstract

Purpose: The risk of developing metastatic squamous cell carcinoma for patients with head and neck squamous cell carcinoma (HNSCC) is very high. Because these patients are often heavy tobacco users, they are also at risk for developing a second primary cancer, with squamous cell carcinoma of the lung (LSCC) being the most common. The distinction between a lung metastasis and a primary LSCC is currently based on certain clinical and histologic criteria, although the accuracy of this approach remains in question.

Experimental Design: Gene expression patterns derived from 28 patients with HNSCC or LSCC from a single center were analyzed using penalized discriminant analysis. Validation was done on previously published data for 134 total subjects from four independent Affymetrix data sets.

Results: We identified a panel of 10 genes (*CXCL13, COL6A2, SFTPB, KRT14, TSPYL5, TMP3, KLK10, MMP1, GAS1*, and *MYH2*) that accurately distinguished these two tumor types. This 10-gene classifier was validated on 122 subjects derived from four independent data sets and an average accuracy of 96% was shown. Gene expression values were validated by quantitative reverse transcription-PCR derived on 12 independent samples (seven HNSCC and five LSCC). The 10-gene classifier was also used to determine the site of origin of 12 lung lesions from patients with prior HNSCC.

Conclusions: The results suggest that penalized discriminant analysis using these 10 genes will be highly accurate in determining the origin of squamous cell carcinomas in the lungs of patients with previous head and neck malignancies.

Patients with head and neck squamous cell carcinoma (HNSCC) are at high risk for the development of metastatic carcinoma in the lung. Studies suggest that 5% to 15% of patients with HNSCC develop lung metastases (1). However, because patients with HNSCC are often heavy tobacco users, they are also at risk for second primary cancers, with squamous cell carcinoma of the lung (LSCC) being the most common (2).

In some cases, the distinction between a lung metastasis and a second primary lung carcinoma can be easily distinguished on clinical grounds. The presence of multiple pulmonary nodules is usually considered evidence of metastatic disease. However, in subjects who present with a solitary lung nodule, the distinction between metastasis and primary carcinoma can be more problematic. Usually, patients with HNSCC who are found to have solitary pulmonary lesions undergo surgery or needle biopsy with pathologic evaluation. If the lung lesion is also of squamous cell histology, the distinction between metastasis and primary LSCC is extremely difficult. Currently, this distinction is made by comparison of histologic grade or by the presence of other premalignant changes in the respiratory epithelium; however, the accuracy of this approach is unclear.

Making the correct diagnosis has practical importance for choice of therapy. Although patients with either a primary LSCC or a solitary HNSCC metastases may be eligible for surgical resection, the choice of surgical procedure and the use of adjuvant therapy is usually different in these situations. Additionally, patients with early-stage LSCC have a significantly better prognosis than patients with metastatic HNSCC.

Authors' Affiliations: [1]University of Pennsylvania Medical Center and [2]The Wistar Institute, Philadelphia, Pennsylvania; [3]Columbia University Medical Center and [4]Memorial Sloan-Kettering Cancer Center, New York, New York; [5]University of Minnesota Medical Center, Minneapolis, Minnesota; [6]University of North Carolina-Chapel Hill, Chapel Hill, North Carolina; and [7]University of Oxford, Oxford, United Kingdom

Received 7/10/06; revised 12/5/06; accepted 2/21/07.

Grant support: Pennsylvania Department of Health (PA DOH Commonwealth Universal Research Enhancement Program), Tobacco Settlement grants ME01-740 (L.C. Showe) and SAP 4100020718 (L.C. Showe, S.M. Albelda), NSF RCN 0090286 (M.K. Showe), National Cancer Institute T32 CA09171, caBIG subcontract 79522CBS10 (M. Nebozhyn), and National Cancer Institute K12 CA076931 (A. Vachani).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

A. Vachani and M. Nebozhyn contributed equally to this work.

Requests for reprints: Louise C. Showe, The Wistar Institute, 3601 Spruce Street, Philadelphia, PA 19104. E-mail: lshowe@wistar.upenn.edu.

doi:10.1158/1078-0432.CCR-06-1670

Recent gene expression studies have shown the potential to classify the origin of human carcinoma cell lines (3) and human tumors (4, 5). We have compared HNSCC and LSCC tumors using gene expression profiling with the goal of identifying a small number of differentially expressed genes that could ultimately prove useful and practical in distinguishing primary lung cancer from HNSCC metastases to the lung. Using a training set/validation set approach, we show that penalized discriminant analysis (PDA) can correctly classify patients with HNSCC and LSCC with high accuracy using a discriminant model with as few as 10 genes. The gene expression results were further validated by quantitative reverse transcription-PCR (QRT-PCR) data derived from 12 independent samples for 19 genes. Our classification algorithm also correctly classified a set of 12 squamous lung lesions of undetermined origin from patients with prior HNSCC.

Materials and Methods

Patient characteristics and tissue acquisition from the University of Pennsylvania. Primary LSCC tumors were obtained from a tissue bank at the Thoracic Oncology Research Laboratory at the University of Pennsylvania. Lung cancer patients in this study presented to the University of Pennsylvania between 1993 and 2000 and underwent a lobectomy for resection of LSCC. These patients had a confirmed pathologic diagnosis of squamous cell carcinoma and had not received any prior cancer therapy. Clinical data was acquired via retrospective chart review. HNSCC patients in this study were obtained from a Head and Neck Carcinoma Tissue Bank and underwent surgical resection at the University of Pennsylvania between 1998 and 2002 (6). Tissue acquisition was approved by the institutional review board at the University of Pennsylvania.

Intraoperative tumor samples were routinely dissected from surrounding normal tissue, but no microdissection was done. H&E staining was done to verify the presence of >70% tumor cells. Samples were immediately frozen in liquid nitrogen before RNA extraction.

RNA preparation, target preparation, and hybridization. RNA was extracted from the tumor specimens as previously described (7). All hybridization protocols were conducted as described in the Affymetrix GeneChip Expression Analysis Technical Manual at the University of Pennsylvania Microarray Core. RNA was hybridized to Affymetrix U133A GeneChips (Affymetrix) using standard conditions in an Affymetrix fluidics station.

External data sources. Gene expression profiling data of HNSCC and LSCC tumor samples were provided by four external sources. The samples were analyzed on two different Affymetrix chips—U133A and U95Av2. U133A data included 41 HNSCC samples from the University of Minnesota (8). U95Av2 data sets included 11 LSCC samples from Columbia University (9, 10), 21 LSCC samples from the Dana-Farber Cancer Institute (11), and 49 samples (18 LSCC, 31 HNSCC) from Memorial Sloan-Kettering Cancer Center (12). U95Av2 data from 12 squamous cell lung lesions from patients with previous HNSCC were also provided by Memorial Sloan-Kettering Cancer Center (12). The Dana-Farber Cancer Institute data is available online.[8] All other published data was kindly provided by investigators at the individual institutions. Patient characteristics and details of tissue acquisition, RNA isolation, and array hybridization have been previously described for these four data sets.

Identifying U95Av2 and U133A common genes. Common genes were linked between the two chip types using Affymetrix probe set identifiers. Probe sets that were common between the two different platforms (U95Av2 versus U133A) were aligned using the "best match" file.[9] This spreadsheet identifies the probe sets from the two platforms that are most similar based on several factors, including target sequence match and percentage identity. A total of 9,530 probe sets were overlapping between U95Av2 and U133A.

Microarray normalization. The CEL files for each data set were reprocessed using a publicly available implementation of Robust Multichip Average expression summary (RMAExpress) Version 0.3 (13). Default settings were used for background adjustment, quantile normalization, and log 2 transformation. Samples from the different institutions were processed as independent groups.

Distance weighted discrimination. The distance weighted discrimination (DWD) method is a generalization of the support vector machine, a multivariate technique (14). DWD has been previously shown to be well suited for correction of the systematic biases associated with micro array data sets (15). DWD performance and robust quantification of systematic bias has been reported to be superior to that of classic methods (such as principal component analysis, linear discriminant analysis, and standard linear support vector machine). A detailed description of DWD is given in ref. (16). The DWD calculations were carried out using a Java-based version of DWD method.[10] The following settings were used for the input variables: (*a*) DWD type, nonstandardized DWD, and (*b*) mean adjustment type, centered at the second mean.

Hierarchical clustering. Hierarchical clustering was done using the Pearson correlation distance metric and Ward's linkage. For visual enhancement of Figs. 1 and 2 (showing the results of biclustering of samples and the selected genes), the clustering was carried out after the values for each gene were converted to z scores by subtracting the corresponding gene mean that was computed over all samples being clustered, and dividing by the corresponding SD. Additionally, to keep figure space as compact as possible, the relative length of the main stem that partitions the clustered samples into two main subclusters has been reduced 5-fold in Fig. 3A and B, depicting clustering of samples before and after DWD transformation.

Selection of biomarkers. Identification of genes that were differentially expressed between HNSCC and LSCC was first carried out by multivariate PDA (17, 18). PDA is an extension of classic Fisher linear discriminant analysis (19) applied to problems where the number of covariates (genes) exceeds the number of observations (samples) in the training set (17, 18).

PDA with recursive feature elimination. The genes that contribute the most to the classification model were selected as follows: ~30% of the genes least differentially expressed between HNSCC and LSCC data sets were first eliminated based on the P values from a univariate t test. A progressive scheme of gene reduction is then applied and the least informative genes (usually from 1% to 10%) are removed iteratively. This process is repeated until only one gene remains. A discriminant model is fitted at each reduction and each gene is assigned a computed "predictive power" (discriminant weight × SD), which estimates the contribution of that gene to the discriminant score. The discriminant scores (either positive or negative) define which of the two experimental classes a particular sample belongs and how well each sample is classified.

Resampling procedure. To evaluate the robustness of our classifier and to estimate the confidence intervals for the classification scores for each sample in the independent validation set, PDA with recursive feature elimination was carried out on 100 subsets of the University of Pennsylvania training set and applied to classify the validation samples. The 100 training subsets were generated by random resampling without replacement (jackknifing) from 28 samples in the University of Pennsylvania data set. Each subset contained 90% of

[8] http://research.dfci.harvard.edu/meyersonlab/lungca/

[9] http://www.affymetrix.com/support/technical/comparison_spreadsheets.affx

[10] https://genome.unc.edu/pubsup/dwd

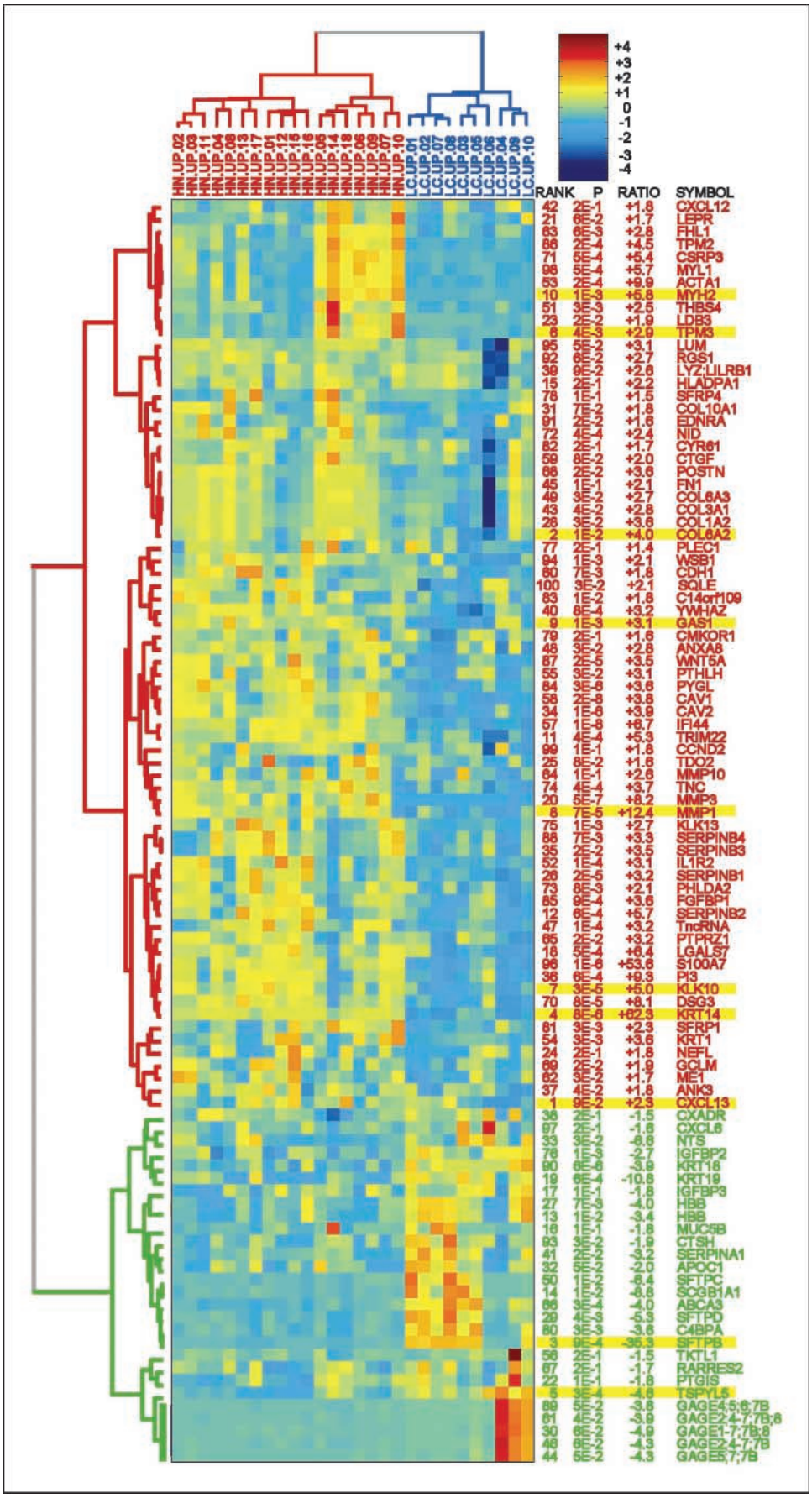

Fig. 1. Hierarchical clustering of the University of Pennsylvania HNSCC and LSCC samples using the top 100 probe sets (representing 99 unique genes) that were identified by PDA with recursive feature elimination as described in Materials and Methods. The PDA was trained to distinguish between 18 HNSCC and 10 primary LSCC samples from the University of Pennsylvania data set (the training set). The genes were selected using the 9,530 probe sets common to U95Av2 and U133A chips. Heat map, rescaled gene expression values for each of the samples and probe sets. For each probe set, the values in the blue range of spectrum correspond to low values and red correspond to high values, relative to the mean expression each probe set, as indicated on the color bar (*upper right corner*). Columns from left to right, gene rank, *P* value, fold change, and gene symbol. Yellow, 10 top-ranked genes.

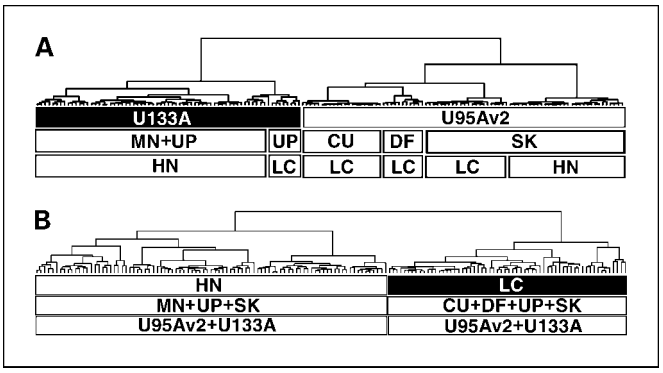

Fig. 2. Unsupervised hierarchical clustering of all 150 samples using the 9,530 overlapping genes representing "perfect match" probe sets common between U95Av2 and U133A chips (analyzed by Ward's linkage and Pearson correlation – based distance metric). *A,* samples clustered first according to Affymetrix chip (U95Av2 versus U133A) used and then according to the source of the data [University of Pennsylvania (*UP*), Dana-Farber Cancer Institute (*DF*), Memorial Sloan-Kettering Cancer Center (*SK*), University of Minnesota (*MN*), Columbia University (*CU*)]. *B,* clustering results after systematic bias adjustment with DWD. Samples now cluster according to tumor type, and no subclustering by Affymetrix chip type is observed.

the 28 original samples, with the same proportion of LSCC and HNSCC.

Quantitative real-time PCR. Gene-specific primers (IDT, Inc.) were designed with the Light Cycler Probe Design Software, Version 1.0 (Idaho Technology, Inc.), and ABI PRISM PrimerExpress software, Version 2.0 (Applied Biosystems, Inc.). Primers were selected from the 3′ half of the message using sequence retrieved from the Genbank database and in almost all cases from different exons. The PCR reaction was done in 20 μL as previously described (20) using the Chromo4 PTC-200 Peltier Thermal Cycler (MJ Research). All primers were designed to have a melting temperature of ~60°C. The PCR cycle variables were as follows: 95°C 3 min hot start, 40 cycles of 95°C 20 s, 60°C 10 s, 72°C 20 s, and 78°C 5 s (to ensure elimination of side product). SYBR green I fluorescence intensity was measured at the end of each 72°C extension as previously described (20). Results were normalized to *GAPDH* as the housekeeping gene and values were calculated relative to a standard curve generated using the Stratagene universal standard RNA, which had been supplemented with RNA from the Jar and HT3 epithelial cell lines. The same standard RNA mixture was used for all comparisons. Product specificity was assessed by melting curve analysis and selected samples were run on 2% agarose gels for size assessment. Quality of real-time PCR was determined in two ways: the amplification efficiencies had to be 100 ± 10%, and correlation coefficients (r^2) >95%. The cDNA for PCR amplification were prepared from 0.5 μg of the amplified RNA using Superscript II as previously described. The amplified RNA was generated from 250 ng of total RNA subjected to one round of linear amplification using the RiboAmp RNA Amplification Kit (Arcturus, Inc.). Some samples were also assayed from cDNA prepared from total RNA with similar results.

Results

Clinical characteristics of the University of Pennsylvania training set cohort. Twenty-eight patients that underwent surgical resection for their primary HNSCC or LSCC were evaluated. The clinical characteristics of these patients are presented in Table 1. In general, the two groups of patients were similar in age, gender, and racial distribution. Clinical data on all 28 subjects was collected via retrospective chart reviews and in certain cases by telephone interview. None of the LSCC patients had a previous history of HNSCC cancer and none developed evidence of HNSCC during a minimum of 5 years of clinical follow-up. Thus, all 10 were judged to have true primary squamous cell carcinoma of the lung.

Training set analysis. We analyzed the 28 patient samples in the University of Pennsylvania training set (18 HNSCC, 10 LSCC) using PDA with recursive feature elimination to identify genes with the highest power to correctly distinguish patients with LSCC from those with HNSCC. Our previous experience had shown that genes selected by PDA did better as classifiers than genes selected by *t* test when applied to a new set of validation samples (20). We trained the PDA program on the 10 LSCC and 18 HNSCC samples from University of Pennsylvania using the 9,530 probe sets representing the overlap between the U95Av2 and U133A arrays. Hierarchical clustering using the top 100 probe sets identified by PDA/recursive feature elimination is shown in Fig. 1. All the samples were correctly separated into the two different phenotypic groups. The top 10 ranked genes are highlighted in yellow. Table 2 lists the 100 genes most significantly differentially expressed between the two patient groups.

Merging the data in the four independent data sets used for validation. We next sought to test the accuracy of the differentially expressed genes identified by PDA on previously published data for independent sets of samples obtained from Memorial Sloan-Kettering Cancer Center, Dana-Farber Cancer Institute, University of Minnesota, and Columbia University. We first evaluated systematic biases in the data sets that might be due to source (where the samples were isolated and processed) or to the array platform used (U95Av2 or U133A). When we tested the University of Pennsylvania data set by hierarchical clustering using the 9,530 overlapping genes, we got a perfect separation by phenotype as expected for these different tumor types (Supplementary Fig. S1). We then applied unsupervised hierarchical clustering using the 9,530 common genes to all 150 samples (including those from University of

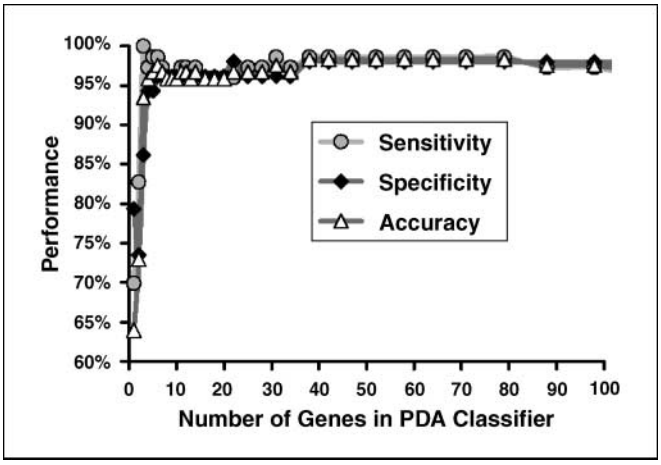

Fig. 3. Accuracy of the PDA classifier on the independent test set after systematic bias correction by DWD. Sensitivity, specificity, and the mean accuracy are given as a function of the number of genes used by the linear discriminant function throughout recursive feature elimination. Further details are given in Materials and Methods. Data are shown for the top 100 genes. For >100 genes, the sensitivity, specificity, and the mean accuracy were essentially the same. There is a small but not significant reduction in accuracy around 40 genes; however, the accuracy is essentially unchanged between 100 and 5 genes.

Table 1. Patient characteristics (University of Pennsylvania subjects)

Variable	HNSCC (n = 18)	LSCC (n = 10)
Age, y, mean (SD)	61	62
Gender (%)		
Male	78	80
Female	22	20
Race (%)		
White	83	90
Black	11	10
Other	6	0
Pathologic T stage (%)		
T_1	11	30
T_2	28	30
T_3	17	10
T_4	44	30
Pathologic N stage (%)		
N_0	39	80
N_1	11	20
N_2	50	0
Histologic grade (%)		
1	5	0
2	56	70
3	39	30
Tumor site (%)		
FOM/buccal/tonsil	22	NA
Gingiva	6	NA
Larynx	0	NA
Mandible	11	NA
Tongue	61	NA

Abbreviations: FOM, floor of mouth; NA, not available.

Pennsylvania). As shown in Fig. 2A, rather than clustering by tumor type, the samples clustered first according to the Affymetrix chip (U95Av2 versus U133A) used for the study and then according to the source of the data (University of Pennsylvania, Dana-Farber Cancer Institute, Memorial Sloan-Kettering Cancer Center, University of Minnesota, and Columbia University). To minimize the artificial variability due to different institutions and chip versions, we applied DWD. DWD is designed to correct the systematic bias in one data set at a time, and in our case several data sets with multiple biases due to the data source and the chip used for hybridization (15). The DWD correction was carried out in the following order but the results were essentially the same regardless of the order in which the data were merged. First, the 41 HNSCC samples in the University of Minnesota data set were merged with the 18 HNSCC samples from University of Pennsylvania (both on U133A). The 11 LSCC samples from Columbia University were then merged with the 10 University of Pennsylvania LSCC samples (U95Av2 and U133A, respectively). The 21 Dana-Farber Cancer Institute LSCC samples were also merged with the 10 LSCC University of Pennsylvania samples. Finally, the two data sets with values for both HNSCC and LSCC, those from Memorial Sloan-Kettering Cancer Center and the University of Pennsylvania, were merged. Hierarchical clustering done after DWD correction is shown in Fig. 2B. All 150 samples now clustered according to their tumor type and no subclustering by chip type or location is observed.

Validation of the discriminant model on the independent test set. The discriminant model using the genes identified by PDA with recursive feature elimination on University of Pennsylvania training set was then applied to classify 72 HNSCC and 50 LSCC samples in the DWD adjusted validation cohort. The observed accuracy of classification as a function of the total number of genes retained in the discriminant model is shown in Fig. 3. Values are shown for classifiers ranging from 1 to 100 genes. There is little change in accuracy between 100 and 5 genes. Because the classification accuracies were essentially the same with 5 or 10 genes, we used the 10 genes in further tests to accommodate greater heterogeneity that may exist in a larger sample set. Using this 10-gene classifier, the measurements of average accuracy, sensitivity, and specificity were each calculated to be 96%. Therefore, 10 genes are sufficient to robustly discriminate the HNSCC samples from LSCC samples in the validation set.

In applying the 10-gene classifier, each sample in the validation set is given a discriminant score that is a measure of how well it is classified. The discriminant scores for each individual subject in the validation cohort are shown in Fig. 4. Of the 122 total samples, only five samples were misclassified, three samples were LSCC, and two samples were HNSCC. Two of the misclassified LSCC samples were considered to be borderline cases. These samples had a low predictive score, shown by the low column height in Fig. 4, and error bars that cross the zero line separating the two classes. The 10 genes used for this classification include chemokine ligand 13 (*CXCL13*); collagen, type VI, α2 (*COL6A2*); surfactant protein B (*SFTPB*); keratin 14 (*KRT14*); TSPY-like 5 (*TSPYL5*); tropomyosin 3 (*TMP3*); kallikrein 10 (*KLK10*); matrix metalloproteinase 1 (*MMP1*); growth arrest-specific 1(*GAS1*); and myosin, heavy polypeptide 2, skeletal muscle, adult (*MYH2*). These are highlighted in yellow on the tree view in Fig. 2.

Validation of selected gene expressions. The gene expression values determined for the University of Pennsylvania array data set were confirmed using two methods. First, the gene expression ratios (HNSCC/LSCC) of 19 genes were compared with the ratios obtained for the same genes in the Memorial Sloan-Kettering Cancer Center data set, the only other data set that included samples from both tumor types. As seen in Table 3, there is a very high level of agreement between the University of Pennsylvania and Memorial Sloan-Kettering Cancer Center data sets. Second, QRT-PCR on samples derived from a new group of seven HNSCC subjects and five LSCC subjects enrolled at University of Pennsylvania and not previously analyzed on microarrays was used to confirm gene expression ratios determined by microarrays on these 19 genes. These new samples provided a second validation set for testing the classifier. Only one of the 19 genes selected had an expression ratio that did not agree among the three data sets—COL6A2 had higher expression in HNSCC compared with LSCC in the University of Pennsylvania array study and by QRT-PCR, whereas its expression in HNSCC was slightly lower in the Memorial Sloan-Kettering Cancer Center data set.

Diagnostic accuracy of gene expression ratios by RTQ-PCR. The QRT-PCR data was used to generate gene expression ratios (nine HNSCC, seven LSCC) using the method of Gordon et al. (21). This method relies on the selection of gene pairs that are highly differentially expressed between the two patient classes. Briefly, the expression values for genes found to be expressed at significantly higher levels in HNSCC were divided by the gene expression values of genes expressed at high levels in LSCC but much lower levels in HNSCC. The diagnostic accuracies of the 10 best performing ratios for 12 samples (nine

Table 2. One hundred most differentially expressed genes (by PDA)

Gene symbol	Gene title	UniGene ID	PDA rank	Mean ratio
CXCL13	*Chemokine (C-X-C motif) ligand 13*	Hs.100431	1	+2.3
COL6A2	*Collagen, type VI, α 2*	Hs.420269	2	+4.0
SFTPB	*Surfactant, pulmonary-associated protein B*	Hs.512690	3	-35.3
KRT14	*Keratin 14*	Hs.355214	4	+62.3
TSPYL5	*TSPY-like 5*	Hs.173094	5	-4.6
TPM3	*Tropomyosin 3*	Hs.146070	6	+2.9
KLK10	*Kallikrein 10*	Hs.275464	7	+5.0
MMP1	*Matrix metalloproteinase 1*	Hs.83169	8	+12.4
GAS1	*Growth arrest-specific 1*	Hs.65029	9	+3.1
MYH2	*Myosin, heavy polypeptide 2, skeletal muscle, adult*	Hs.513941	10	+5.8
TRIM22	*Tripartite motif-containing 22*	Hs.501778	11	+5.3
SERPINB2	*Serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 2*	Hs.514913	12	+5.7
HBB	*Hemoglobin, β*	Hs.523443	13	-3.4
SCGB1A1	*Secretoglobin, family 1A, member 1 (uteroglobin)*	Hs.523732	14	-8.8
HLA-DPA1	*MHC, class II, DP α 1*	Hs.347270	15	+2.2
MUC5B	*Mucin 5, subtype B, tracheobronchial*	Hs.523395	16	-1.8
IGFBP3	*Insulin-like growth factor binding protein 3*	Hs.450230	17	-1.8
LGALS7	*Lectin, galactoside-binding, soluble, 7 (galectin 7)*	Hs.99923	18	+6.4
KRT19	*Keratin 19*	Hs.514167	19	-10.8
MMP3	*Matrix metalloproteinase 3*	Hs.375129	20	+8.2
LEPR	*Leptin receptor*	Hs.23581	21	+1.7
PTGIS	*Prostaglandin I2 (prostacyclin) synthase*	Hs.302085	22	-1.8
LDB3	*LIM domain binding 3*	Hs.49998	23	+1.9
NEFL	*Neurofilament, light polypeptide 68 kDa*	Hs.521461	24	+1.8
TDO2	*Tryptophan 2,3-dioxygenase*	Hs.183671	25	+1.6
SERPINB1	*Serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 1*	Hs.381167	26	+3.2
HBB	*Hemoglobin, β*	Hs.523443	27	-4.0
COL1A2	*Collagen, type I, α2*	Hs.489142	28	+3.6
SFTPD	*Surfactant, pulmonary-associated protein D*	Hs.253495	29	-5.3
GAGE1	*G antigen 1*	Hs.278606	30	-4.9
COL10A1	*Collagen, type X, α1 (Schmid metaphyseal chondrodysplasia)*	Hs.520339	31	+1.8
APOC1	*Apolipoprotein C-I*	Hs.110675	32	-2.0
NTS	*Neurotensin*	Hs.80962	33	-6.6
CAV2	*Caveolin 2*	Hs.212332	34	+3.9
SERPINB3	*Serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 3*	Hs.227948	35	+3.5
PI3	*Protease inhibitor 3, skin-derived (SKALP)*	Hs.112341	36	+9.3
ANK3	*Ankyrin 3, node of Ranvier (ankyrin G)*	Hs.499725	37	+1.8
CXADR	*Coxsackie virus and adenovirus receptor*	Hs.473417	38	-1.5
LYZ	*Lysozyme (renal amyloidosis)*	Hs.524579	39	+2.6
YWHAZ	*Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, ζ polypeptide*	Hs.492407	40	+3.2
SERPINA1	*Serine (or cysteine) proteinase inhibitor, clade A (α-1 antiproteinase, antitrypsin), member 1*	Hs.525557	41	-3.2
CXCL12	*Chemokine (C-X-C motif) ligand 12*	Hs.522891	42	+1.8
COL3A1	*Collagen, type III, α1*	Hs.443625	43	+2.8
GAGE5	*G antigen 5*	Hs.278606	44	-4.3
FN1	*Fibronectin 1*	Hs.203717	45	+2.1
GAGE2	*G antigen 2*	Hs.278606	46	-4.3
TncRNA	*Trophoblast-derived noncoding RNA*	Hs.523789	47	+3.2
ANXA8	*Annexin A8*	Hs.463110	48	+2.8

NOTE: PDA rank is the order in which the given gene was eliminated during the course of recursive feature elimination. Mean ratio is ratio of mean gene expression ratio in one group versus the other. A positive ratio corresponds to higher expression in HNSCC and a negative ratio corresponds to higher expression in LSCC.

HNSCC, seven LSCC) are presented in Fig. 5. All nine of these gene ratios accurately separated the two tumor types with differences approaching 1,000-fold in some cases.

Classification of lung squamous cell tumors of undetermined origin in patients with previous HNSCC. Having identified a 10-gene classifier with high accuracy for distinguishing between primary lung carcinoma from head and neck carcinomas, we then asked whether our algorithm would be similarly accurate in the classification of 12 squamous cell lung tumors resected from 12 patients previously treated for primary HNSCC. Classification of these samples using a 500-gene classifier that discriminated HNSCC from LSCC was reported previously (12). Although these samples could not definitively be distinguished as lung primaries or lung metastases, based on pathologic and clinical criteria most of the lesions were suspected to be of lung origin (12). When our 10-gene classifier

Table 2. One hundred most differentially expressed genes (by PDA) (Cont'd)

Gene symbol	Gene title	UniGene ID	PDA rank	Mean ratio
COL6A3	*Collagen, type VI, α3*	Hs.233240	49	+2.7
SFTPC	*Surfactant, pulmonary-associated protein C*	Hs.1074	50	-6.4
THBS4	*Thrombospondin 4*	Hs.211426	51	+2.5
IL1R2	*Interleukin 1 receptor, type II*	Hs.25333	52	+3.1
ACTA1	*Actin, α1, skeletal muscle*	Hs.1288	53	+9.9
KRT1	*Keratin 1 (epidermolytic hyperkeratosis)*	Hs.80828	54	+3.6
PTHLH	*Parathyroid hormone-like hormone*	Hs.89626	55	+3.1
TKTL1	*Transketolase-like 1*	Hs.102866	56	-1.5
IFI44	*IFN-induced protein 44*	Hs.82316	57	+6.7
CAV1	*Caveolin 1, caveolae protein, 22 kDa*	Hs.74034	58	+3.8
CTGF	*Connective tissue growth factor*	Hs.410037	59	+2.0
CDH1	*Cadherin 1, type 1, E-cadherin (epithelial)*	Hs.461086	60	+1.8
GAGE2	*G antigen 2*	Hs.278606	61	-3.9
ME1	*Malic enzyme 1, NADP(+)-dependent, cytosolic*	Hs.21160	62	+1.7
FHL1	*Four and a half LIM domains 1*	Hs.435369	63	+2.8
MMP10	*Matrix metalloproteinase 10*	Hs.2258	64	+2.6
PTPRZ1	*Protein tyrosine phosphatase, receptor-type, Z polypeptide 1*	Hs.489824	65	+3.2
ABCA3	*ATP-binding cassette, subfamily A (ABC1), member 3*	Hs.26630	66	-4.0
RARRES2	*Retinoic acid receptor responder (tazarotene induced) 2*	Hs.521286	67	-1.7
POSTN	*Periostin, osteoblast specific factor*	Hs.136348	68	+3.6
GCLM	*Glutamate-cysteine ligase, modifier subunit*	Hs.315562	69	+1.9
DSG3	*Desmoglein 3 (pemphigus vulgaris antigen)*	Hs.1925	70	+8.1
CSRP3	*Cysteine and glycine-rich protein 3 (cardiac LIM protein)*	Hs.83577	71	+5.4
NID	*Nidogen (enactin)*	Hs.356624	72	+2.4
PHLDA2	*Pleckstrin homology-like domain, family A, member 2*	Hs.154036	73	+2.1
TNC	*Tenascin C (hexabrachion)*	Hs.143250	74	+3.7
KLK13	*Kallikrein 13*	Hs.165296	75	+2.7
IGFBP2	*Insulin-like growth factor binding protein 2, 36 kDa*	Hs.438102	76	-2.7
PLEC1	*Plectin 1, intermediate filament binding protein 500 kDa*	Hs.434248	77	+1.4
SFRP4	*Secreted frizzled-related protein 4*	Hs.105700	78	+1.5
CMKOR1	*chemokine orphan receptor 1*	Hs.471751	79	+1.6
C4BPA	*Complement component 4 binding protein, α*	Hs.1012	80	-3.6
SFRP1	*Secreted frizzled-related protein 1*	Hs.213424	81	+2.3
CYR61	*Cysteine-rich, angiogenic inducer, 61*	Hs.8867	82	+1.7
C14orf109	*Chromosome 14 open reading frame 109*	Hs.275352	83	+1.8
PYGL	*Phosphorylase, glycogen; liver (Hers disease, glycogen storage disease type VI)*	Hs.282417	84	+3.6
FGFBP1	*Fibroblast growth factor binding protein 1*	Hs.1690	85	+3.6
TPM2	*Tropomyosin 2 (β)*	Hs.300772	86	+4.5
WNT5A	*Wingless-type MMTV integration site family, member 5A*	Hs.152213	87	+3.5
SERPINB4	*Serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 4*	Hs.123035	88	+3.3
GAGE4	*G antigen 4*	Hs.278606	89	-3.8
KRT18	*Keratin 18*	Hs.406013	90	-3.9
EDNRA	*Endothelin receptor type A*	Hs.183713	91	+1.6
RGS1	*Regulator of G-protein signaling 1*	Hs.75256	92	+2.7
CTSH	*Cathepsin H*	Hs.148641	93	-1.9
WSB1	*WD repeat and SOCS box-containing 1*	Hs.446017	94	+2.1
LUM	*Lumican*	Hs.406475	95	+3.1
S100A7	*S100 calcium binding protein A7 (psoriasin 1)*	Hs.112408	96	+53.6
CXCL6	*Chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2)*	Hs.164021	97	-1.6
MYL1	*Myosin, light polypeptide 1, alkali; skeletal, fast*	Hs.187338	98	+5.7
CCND2	*Cyclin D2*	Hs.376071	99	+1.8
SQLE	*Squalene epoxidase*	Hs.71465	100	+2.1

was tested on these 12 samples, the majority (U01 through U11) had strong negative classification scores, identifying them as primary lung carcinomas as was suspected (Fig. 6; ref. 12). Sample U12, which was clinically thought to be a lung metastasis based on the development of an additional pancreatic metastasis in the patient, is classified as HNSCC supporting the clinical impression. Sample U13, which is the pancreatic metastasis from the same patient as U12, is also classified as HNSCC, although not as robustly, suggesting it retains the HNSCC gene signature.

Discussion

Relevance. The finding of a solitary pulmonary lesion of squamous cell histology in a patient with HNSCC can represent either a metastasis, or more likely, a primary lung tumor (22).

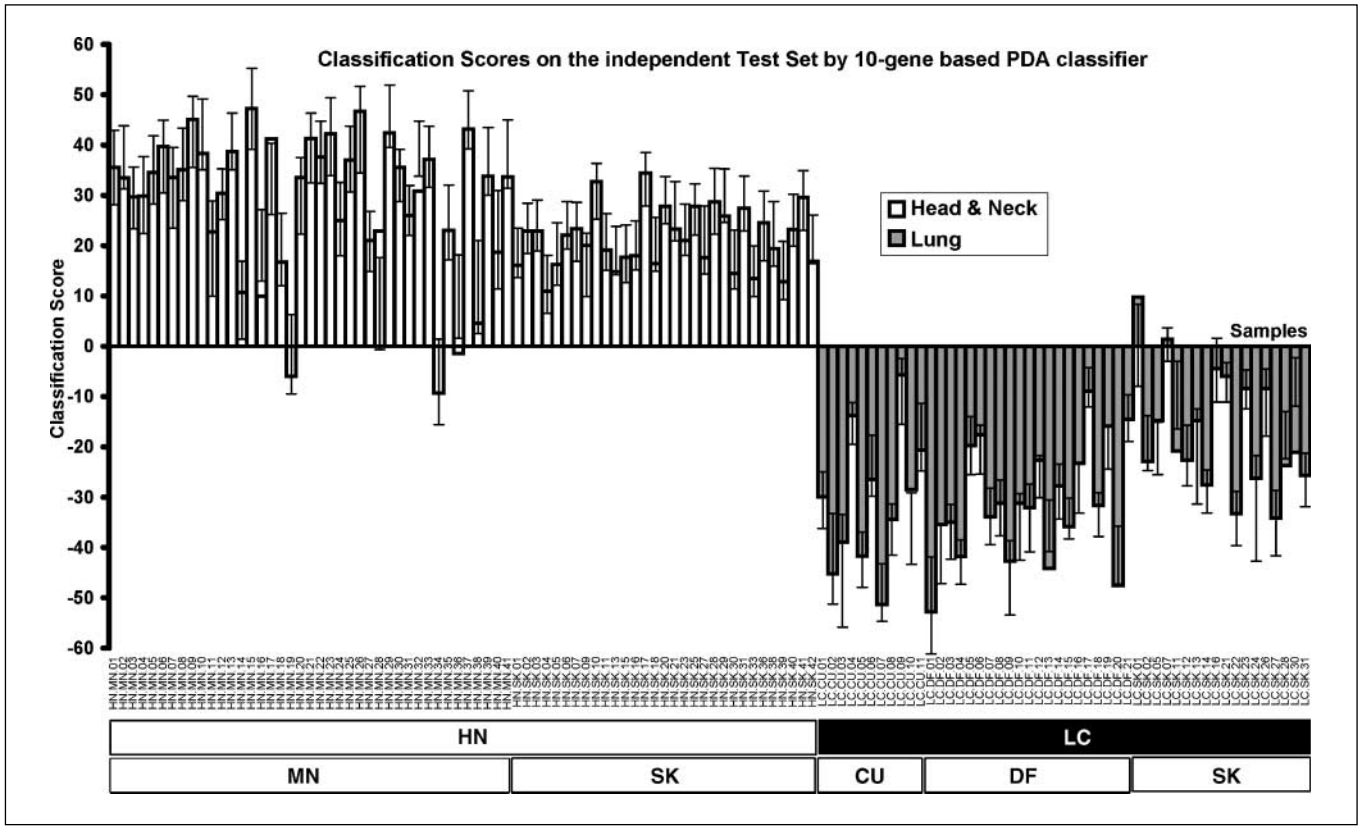

Fig. 4. Discriminant scores assigned to 122 patient samples in the validation set. A positive score corresponds to classification of a sample as HNSCC; a negative score indicates classification as LSCC. Columns, classification score for each sample; bars, confidence intervals, centered on the mean and equal to 2 SDs computed over the scores assigned by PDA with recursive feature elimination derived from 100 resamplings of the training set. Samples are arranged by source.

Currently, various clinical criteria are used to distinguish between these two entities. When the two lesions are of similar histologic appearance, the lung nodule is assumed to be metastatic. The presence of malignant adenopathy in the anterior triangle of the neck at the time of diagnosis of the lung lesion also suggests pulmonary metastases. Finally, the presence of a lung lesion within 3 years of the HNSCC also makes metastases more likely.

It is important to differentiate between these possibilities because the prognosis and treatment of a primary versus metastatic lesion are different. Studies have shown that the surgical approach for a primary LSCC should be a lobectomy compared with a lesser resection (23), whereas the goal of resection in pulmonary metastases is to remove all gross tumor while preserving as much normal parenchyma as possible. This can usually be achieved via a wedge resection. Additionally, the role of lymph node dissection, which is standard for primary lung cancer resection, is not well defined in pulmonary metastasectomy (24). The choice of adjuvant therapy is also affected—platinum-based chemotherapy is now frequently used for primary lung cancer, whereas its role after metastasectomy has not been studied. Finally, the 5-year survival of early-stage lung cancer after lobectomy approaches 80%, but is much lower in patients with metastatic disease (25).

Recent studies suggest that the use of genetic abnormalities can help with the distinction between primary LSCC and metastasis. Leong et al. (26) compared tumors from 16 patients with HNSCC and a paired solitary lung nodule for loss of heterozygosity on chromosomal arms 3p and 9p. The use of loss of heterozygosity distinguished 13 of the 16 cases as primary lung cancer or metastasis based on discordant versus concordant allelic patterns between the index tumor and the lung lesion. Of the top 100 genes in our study, only two (*WNT5A*, *TPM2*) are located on one of these chromosomal arms. This is not surprising as both 3p and 9p are frequently lost in both HNSCC and LSCC and would therefore be less likely to lead to identification of differentially expressed genes in these two tumor types.

A separate study using loss of heterozygosity suggests that many squamous lung lesions in patients with HNSCC that are currently classified as metastases based on clinical criteria may in fact be primary lung cancers (27). Although loss of heterozygosity is potentially useful, this technique is time consuming, not widely available, not completely accurate, and, most importantly, requires appropriate tissue from both the primary and the lung lesion.

Comparison with other studies. Although a number of studies have been published examining gene profiles in HNSCC (8, 28) and LSCC (9) with their tissues of origin, to our knowledge, the patterns in these two types of tumors have only been compared in one previous study (12). Talbot et al. used gene expression profiling to compare 21 lung cancer and 31 tongue cancer samples and were able to distinguish between HNSCC and LSCC tumors using hierarchical clustering with 100 to 500 genes. The accuracy of their predictions decreased when the number of genes was reduced below 100. An

important advantage of our discriminant model over the traditional hierarchical clustering/*t* test approach is the accuracy that was achieved using a small number of genes. Our 10-gene classifier also correctly classified 96% of the samples from the Talbot et al. study. Although as few as five genes could be used with equal accuracy, we used the 10-gene classifier in these studies.

A major concern in small array-based studies is the high degree of heterogeneity that exists within a single tumor type and whether the samples properly capture that case to case heterogeneity. In this particular study, factors that have not been considered include tobacco use and human papillomavirus status (for the HNSCC cases). Nevertheless, gene expression differences between the two tumor types were striking and our 10-gene classifier was evaluated with testing and validation data sets using several different groups of external samples. This allowed us to show that the data from University of Pennsylvania used for model building and gene selection was highly accurate when evaluated on these external data sets. Most biomarker studies are done in a single institution and usually on conservative sample numbers. If validation is done, it is usually with split sample or 10-fold cross-validation approaches, which, if not used carefully, can lead to bias in gene selection and "overfitting" of the data (29). The possibility of combining data sets of different origins to avoid this problem is shown by these studies.

Table 3. QRT-PCR expression ratios of selected genes

Gene	QRT-PCR		Penn data set		SK data set	
	7 HNSCC-5 LSCC		18 HNSCC-10 LSCC		31 HNSCC-21 LSCC	
	Ratio	*P*	Ratio	*P*	Ratio	*P*
c1orf42	+466.4	6E-2	+5.0	3E-5	NA	NA
ACTA1	+462.9	3E-1	+9.9	2E-4	+7.5	2E-7
ARNTL2	+8.4	3E-2	+2.9	5E-4	NA	NA
CALD1	+14.9	2E-2	+4.6	2E-4	+1.6	5E-5
CD44	+4.8	2E-1	+3.0	2E-3	+1.3	8E-3
COL6A2	+35.5	3E-2	+4.0	1E-2	-1.1	4E-1
COL6A3	+18.4	4E-2	+2.7	3E-2	+1.2	3E-1
KRT14	+188.9	1E-2	+62.3	8E-6	+29.7	1E-6
KRT16	+14.8	9E-2	+15.8	6E-5	+9.7	9E-7
KRT17	+29.0	3E-2	+10.6	9E-4	+2.3	7E-4
MMP1	+56.2	1E-2	+12.4	7E-5	+6.2	2E-6
S100A7	+357.9	2E-2	+53.6	7E-7	+35.2	8E-9
TPM2	+12.1	1E-1	+4.5	2E-4	+2.0	2E-4
TRAIL	+3.7	7E-1	+4.2	3E-3	+1.9	3E-3
SCGB1A1	-113.6	3E-2	-8.8	1E-2	-1.8	2E-3
SFTPA2	-10,245.4	4E-3	-19.5	1E-3	NA	NA
SFTPB	-3,920.5	4E-3	-35.3	9E-4	-12.4	1E-5
SFTPC	-5.0	2E-1	-6.4	1E-2	-4.5	2E-4
SFTPD	-16.6	6E-2	-5.3	4E-3	-1.9	1E-4

NOTE: Comparison of the results for the *P* values and the ratios of 19 genes found to be differentially expressed between HNSCC and LSCC on University of Pennsylvania data set, as validated by microarrays on the independent set from Memorial Sloan-Kettering Cancer Center, and confirmed by QRT-PCR on a separate set of samples. *P* values for QRT-PCR data were estimated by the Mann-Whitney test.

Abbreviations: Penn, University of Pennsylvania; SK, Memorial Sloan-Kettering Cancer Center.

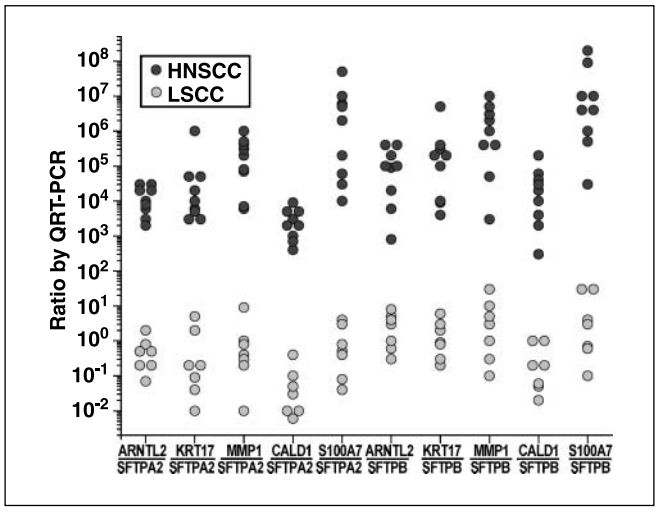

Fig. 5. Ten selected ratios for five up-regulated (*ARNTL2, CALD1, KRT17, MMP1, S100A7*) and two down-regulated genes (*SFTPA2, SFTPB*) for nine HNSCC and seven LSCC samples using amplified RNA for QRT-PCR. The numerical values for the ratios of two genes in the same sample are shown relative to that of the reference Stratagene control.

Analysis of specific pathways and genes. Some of the most useful discriminating genes we detected were the lung surfactant genes, which were significantly higher in the LSCC. This is not surprising, given the lung epithelial origin of these tumors. However, because the tumor samples used for the gene expression studies were not microdissected and thus potentially contained up to 30% nontumor tissues, a potential explanation for this finding of high surfactant gene expression in the LSCC, but not in HNSCC, could be contamination from normal lung tissue in our original LSCC samples. Because of the availability of an antisurfactant protein C polyclonal antibody that worked well in paraffin-fixed tissues, we stained some of the LSCC specimens to determine the cellular localization of the SP-C. Our staining studies showed strong cytoplasmic staining in LSCC tumor cells (data not shown), demonstrating that the increased gene expression was not simply due to contaminating lung tissues. In addition, the 10-gene classifier, including the surfactant genes, easily distinguished LSCC from normal adjacent lung tissue using data available in the Memorial Sloan-Kettering Cancer Center data set further supporting the observation that the differential expression is tumor associated (data not shown).

Another major gene family with increased expression in lung cancers is the GAGE (G antigen) genes. GAGE proteins are a large group of cancer/testis antigens consisting of GAGE-1 through GAGE-8 (30). Although the function of most of the cancer/testis antigens is not known, GAGE proteins have been implicated in inhibition of apoptosis and chemotherapy resistance (31, 32). GAGE protein expression is present in ~40% of lung cancers and is associated with poor prognosis (33). Detrimental effects of GAGE expression on survival has also been shown in esophageal and brain tumors (34, 35). Interestingly, GAGE gene expression was up-regulated in only a subset of LSCC (and no HNSCC). The significance of these proteins in the pathophysiology or prognosis of these tumors is as yet unknown.

One of the most striking changes we observed in our data set was difference in expression of specific cytokeratin genes in

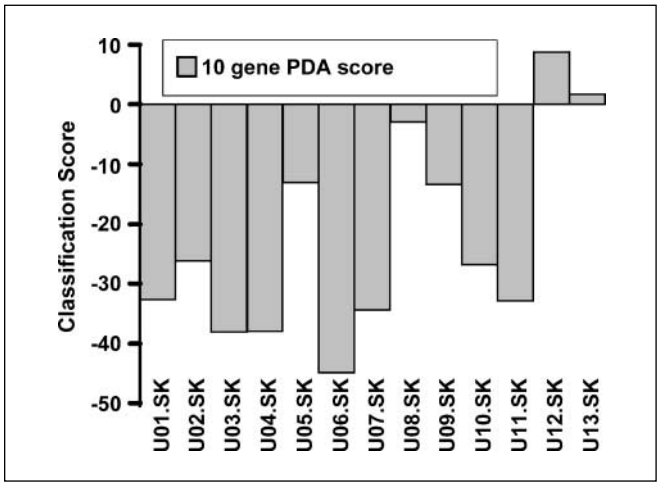

Fig. 6. Classification scores assigned to the 12 LSCC tumors and 1 pancreatic lesion derived from patients with previous HNSCC described in Talbot et al. (12). Positive scores correspond to samples classified as HNSCC and negative score indicates classification as LSCC. Columns, scores generated by the 10-gene PDA classifier. The results were obtained using gene expression data for the 13 samples in the test set without any systematic bias adjustment.

these two types of tumors. All eukaryotic cells contain a cytoskeleton composed of three distinct filamentous structures: microfilaments, intermediate filaments, and microtubules (36). The intermediate filament protein family includes several hundred different members that are divided into several groups. Cytokeratins constitute type I and type II intermediate filaments and are subdivided based on isoelectric point (CKs 1-9 are acidic; CKs 10-20 are basic). Stratified squamous epithelia express mostly CKs 1 to 6 and 9 to 17, whereas CKs 7, 8, and 18 to 20 are identified in simple epithelia (36). During malignant transformation of normal cells, the cytokeratin patterns are usually maintained.

The pattern of gene expression differences identified in our study showed a "stratified squamous epithelial" pattern in the HNSCC tumors with higher expression of CKs 1 and 14 (up 3.6- and 62-fold, respectively) and lower expression of CKs 18 and 19 (down 3.9- and 10-fold, respectively). Although both upper airway epithelium and bronchial epithelium are composed of stratified squamous cells, it is not surprising that HNSCC tumors are more likely to exhibit a stratified squamous pattern given their location in the upper aerodigestive tract.

Many genes in the collagen family were also up-regulated in head and neck tumors when compared with squamous cell lung cancer. Five collagen-related genes (*COL6A2*, *COL1A2*, *COL10A1*, *COL3A1*, and *COL6A3*) were found in our top 100 genes selected by PDA and had expression ratios ranging from +1.8 to +4.0. In the tumor microenvironment, collagens are a major component of the extracellular matrix, which is primarily secreted by stromal cells and inflammatory cells (37). Thus, the higher expression of collagen in the head and neck tumors may simply reflect a higher proportion of stromal elements compared with the lung cancer samples. There is recent data, however, that suggests that certain collagen genes are expressed in the tumor cells themselves. For example, ovarian cancer cells have been shown to highly express several extracellular matrix proteins, including collagen VI, and this was associated with resistance to cisplatin *in vitro* (38).

The high expression of collagens in the head and neck tumors was mirrored by higher levels of three matrix metalloproteinases, MMP1, MMP3, and MMP10, which were increased by 12.4-, 8.2-, and 2.6-fold, respectively, when compared with the lung cancers. MMP-1, or collagenase-1, is expressed in a wide variety of cancers and in most cases is associated with increased invasion and poorer survival (39). MMP-3, which is a secreted by fibroblasts, can activate tumor-derived MMP-1 and other collagenases leading to increased collagen degradation and tumor invasion (39). In head and neck tumors, high levels of MMP-1 and MMP-3 are associated with greater tumor invasiveness and incidence of lymph node metastases (40). The higher levels of MMP gene expression in our study may have been due to higher proportion of HNSCC tumors with lymph node metastases when compared with the LSCC tumors (61% versus 20%).

Significance and future directions. Although our data was derived from primary LSCC and HNSCC samples, we postulate that our predictive approach will be able to determine the origin of lung nodules in patients with previous HNSCC. We were able to conduct a first test of this hypothesis by validating our 10-gene classifier using data provided by Talbot et al. (12) from a set of 12 squamous cell lung lesions of "unknown etiology" derived from patients with previous HNSCC. As shown in Fig. 6, our predictions closely matched the results of the 500-gene classifier set of Talbot and were consistent with the final clinical classifications of these tumors (12). How the 10-gene classifier performs in a prospective series of squamous cell lung nodules from patients with HNSCC is the subject of ongoing investigation.

In addition to using microarray data, we are also studying ways to use our data in other types of assays. We have, in a limited fashion, shown the use of gene expression ratios using QRT-PCR to distinguish between these two tumor types. This method, originally developed by Gordon et al., has several potential advantages: It does not require a housekeeping gene to be used as a reference, is independent of the platform used for data acquisition, and requires very small amounts of RNA. We ultimately plan to develop PCR classifiers that can be used in paraffin-embedded tissues. Recent advances in PCR technology allow the measurement of gene expression from RNA harvested from paraffin-embedded tumors, which are most commonly used for standard clinical pathology (41). We are currently developing and testing PCR primers that will work well in paraffin-fixed tissue and are collecting a series of well-characterized pathologic specimens to validate this approach. We are also evaluating potential immunohistochemical markers, such as antisurfactant protein C antibodies using tissue arrays. The use of antibody staining remains the most commonly used technique in diagnostic pathology and would be the method most easily adopted into routine clinical practice. If protein ratios mirror the RNA ratios, this could be a useful diagnostic approach.

Summary. The ongoing refinements in surgical therapy and in adjuvant chemotherapy for head and neck cancer and lung cancer make the distinction between primary LSCC and lung metastasis increasingly important. We have identified a 10-gene classifier that we believe can distinguish primary squamous cell tumors of each type. This finding represents a potentially

exciting new molecular diagnostic method, but will need to be further validated before it can be used clinically. We are now actively pursuing the use of both gene expression and immunohistochemical methods in HNSCC patients who present with a solitary lung nodule to further validate our result. Because there is not yet a true "gold standard," our assessment of accuracy validation will require careful and somewhat long-term clinical follow-up.

Acknowledgments

We thank Dr. Michael Feldman for his assistance with this project.

References

1. Ferlito A, Shaha AR, Silver CE, et al. Incidence and sites of distant metastases from head and neck cancer. ORL J Otorhinolaryngol Relat Spec 2001;63:202–7.
2. Jones AS, Morar P, Phillips DE, et al. Second primary tumors in patients with head and neck squamous cell carcinoma. Cancer 1995;75:1343–53.
3. Nishizuka S, Chen ST, Gwadry FG, et al. Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic, and tissue array profiling. Cancer Res 2003;63:5243–50.
4. Giordano TJ, Shedden KA, Schwartz DR, et al. Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles. Am J Pathol 2001;159:1231–8.
5. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001;98:15149–54.
6. O'Donnell RK, Kupferman M, Wei SJ, et al. Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity. Oncogene 2005;24:1244–51.
7. Singhal S, Wiewrodt R, Malden LD, et al. Gene expression profiling of malignant mesothelioma. Clin Cancer Res 2003;9:3080–97.
8. Ginos MA, Page GP, Michalowicz BS, et al. Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. Cancer Res 2004;64:55–63.
9. Borczuk AC, Gorenstein L, Walter KL, et al. Non-small-cell lung cancer molecular signatures recapitulate lung developmental pathways. Am J Pathol 2003;163:1949–60.
10. Borczuk AC, Shah L, Pearson GD, et al. Molecular signatures in biopsy specimens of lung cancer. Am J Respir Crit Care Med 2004;170:167–74.
11. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 2001;98:13790–5.
12. Talbot SG, Estilo C, Maghami E, et al. Gene expression profiling allows distinction between primary and metastatic squamous cell carcinomas in the lung. Cancer Res 2005;65:3063–71.
13. Bolstad BM, Irizarry RA, Astrand M, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185–93.
14. Vapnik V, Chapelle O. Bounds on error expectation for support vector machines. Neural Comput 2000;12:2013–36.
15. Benito M, Parker J, Du Q, et al. Adjustment of systematic microarray data biases. Bioinformatics 2004;20:105–14.
16. Marron JS, Todd MJ. Distance weighted discrimination technical report no. 1339. Ithaca (NY): School of Operations Research and Industrial Engineering, Cornell University; 2002.
17. Raychaudhuri S. Penalized discriminant analysis. Trends Biochem Sci 2001;19:189–93.
18. Hastie T, Buja A, Tibshirani R. Penalized discriminant analysis. Ann Surg 1995;23:73–102.
19. Fisher RA. The statistical utilization of multiple measurements. Ann Eugen 1938;8:376–86.
20. Kari L, Loboda A, Nebozhyn M, et al. Classification and prediction of survival in patients with the leukemic phase of cutaneous T cell lymphoma. J Exp Med 2003;197:1477–88.
21. Gordon GJ, Jensen RV, Hsiao LL, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 2002;62:4963–7.
22. Cahan WG, Shah JP, Castro EB. Benign solitary lung lesions in patients with cancer. Ann Surg 1978;187:241–4.
23. Ginsberg RJ, Rubinstein LV; Lung Cancer Study Group. Randomized trial of lobectomy versus limited resection for T1 N0 non-small cell lung cancer. Ann Thorac Surg 1995;60:615–22; discussion 22–3.
24. Rusch VW. Pulmonary metastasectomy. Current indications. Chest 1995;107:322–31S.
25. Liu D, Labow DM, Dang N, et al. Pulmonary metastasectomy for head and neck cancers. Ann Surg Oncol 1999;6:572–8.
26. Leong PP, Rezai B, Koch WM, et al. Distinguishing second primary tumors from lung metastases in patients with head and neck squamous cell carcinoma. J Natl Cancer Inst 1998;90:972–7.
27. Geurts TW, Nederlof PM, van den Brekel MW, et al. Pulmonary squamous cell carcinoma following head and neck squamous cell carcinoma: metastasis or second primary? Clin Cancer Res 2005;11:6608–14.
28. Belbin TJ, Singh B, Barber I, et al. Molecular classification of head and neck squamous cell carcinoma using cDNA microarrays. Cancer Res 2002;62:1184–90.
29. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 2002;99:6562–6.
30. Emens LA, Jaffee EM. To live or not to live: that depends on GAGE? Cancer Biol Ther 2002;1:388–90.
31. Cilensek ZM, Yehiely F, Kular RK, et al. A member of the GAGE family of tumor antigens is an anti-apoptotic gene that confers resistance to Fas/CD95/APO-1, Interferon-γ, Taxol and γ-irradiation. Cancer Biol Ther 2002;1:380–7.
32. Duan Z, Duan Y, Lamendola DE, et al. Overexpression of MAGE/GAGE genes in paclitaxel/doxorubicin-resistant human cancer cell lines. Clin Cancer Res 2003;9:2778–85.
33. Melloni G, Ferreri AJ, Russo V, et al. Prognostic significance of cancer-testis gene expression in resected non-small cell lung cancer patients. Oncol Rep 2004;12:145–51.
34. Cheung IY, Chi SN, Cheung NK. Prognostic significance of GAGE detection in bone marrows on survival of patients with metastatic neuroblastoma. Med Pediatr Oncol 2000;35:632–4.
35. Zambon A, Mandruzzato S, Parenti A, et al. MAGE, BAGE, and GAGE gene expression in patients with esophageal squamous cell carcinoma and adenocarcinoma of the gastric cardia. Cancer 2001;91:1882–8.
36. Barak V, Goike H, Panaretakis KW, et al. Clinical utility of cytokeratins as tumor markers. Clin Biochem 2004;37:529–40.
37. Bhowmick NA, Moses HL. Tumor-stroma interactions. Curr Opin Genet Dev 2005;15:97–101.
38. Sherman-Baust CA, Weeraratna AT, Rangel LB, et al. Remodeling of the extracellular matrix through overexpression of collagen VI contributes to cisplatin resistance in ovarian cancer cells. Cancer Cell 2003;3:377–86.
39. Brinckerhoff CE, Rutter JL, Benbow U. Interstitial collagenases as markers of tumor progression. Clin Cancer Res 2000;6:4823–30.
40. Kurahara S, Shinohara M, Ikebe T, et al. Expression of MMPS, MT-MMP, and TIMPs in squamous cell carcinoma of the oral cavity: correlations with tumor invasion and metastasis. Head Neck 1999;21:627–38.
41. Ramaswamy S. Translating cancer genomics into clinical oncology. N Engl J Med 2004;350:1814–6.